What to do today (Tue Jan 12, 2023)?

Part I. Introduction

Part II. Epidemiologic Concepts and Designs

Part II.1 Epidemiologic view of diseases and populations

Part II.2 Measuring disease frequency in population

II.2.2 Disease Frequency: Basics

II.2.3 Disease Frequency: Inferential

Part II.3 Overview of Designs for Medical Studies

II.3.1 Introduction

II.3.2 Types of Study Design

II.3.3. Related Issues

Part II.2.3 Disease Frequency: Inferential (Advanced)

What if limited information available?

$\Longrightarrow$ statistical methods come to work ...

Part II.2.3A Confidence Limits for Disease Prevalence: to estimate the true disease prevalence p when a random sample of size n available from the population

- 95% CI for p: $\hat{p} \pm 1.96 SE(\hat{p})$
- Frequentist and Bayesian interpretations for CI
- Efficiency?
- (i) Effect of n? (ii) Level of confidence?
- More on $\hat{p} \pm 1.96 SE(\hat{p})$, a Wald type CI

Moreover, ...

- when n is small?

- when the size of the defined population is not too large:
  infinite population vs finite population

What if

- the sample is not random?

Confidence intervals account for variation in sampling (random) error (not others!) by the margin of error term

Inferences with data about binary variables?

Part II.2.3B Estimating Cumulative Incidence, Allowing for Censoring:

The **sample (observed) risk** by time t:

- directly calculable for a closed population

$$\textit{Cumulative Incidence} = \frac{\#\text{disease-onset by time t}}{\#\text{people initially at risk}}$$

  - Cumulative incidence by time t is a proportion.
    Methods used to construct CI for proportions may be used when cumulative incidence by different t is directly estimated
- how about for an open population?
- how about for a closed poulation with censoring?

What is censoring?

- type of incomplete data mechanism
  - type 1 censoring
  - type 2 censoring
  - a general right-censoring

observations subject to (right-)censoring

$\Longrightarrow$ (right-)censored data

- How to estimate the survivor function of the disease-onset time T, $S(t) = Pr(T \geq t)$, with censored data?
- Recall Cumulative incidence at time t $= Pr(T \leq t) = 1 - S(t)$

Kaplan-Meier (product-limit) estimator: a nonparametric MLE of $S(t)$ with right-censored times

- a step function of t with steps at the observed disease-onset times

- in the absence of censoring

$$\hat{S}_{KM}(t) = 1 - \frac{\#\textit{events by time } t}{\#\textit{persons initially at risk}}.$$

- Data $\{y_i = (u_i, \delta_i) : i = 1, \ldots n\}$: $u_i = \min(t_i, c_i)$ and $\delta_i = 1, 0$ if $u_i = t_i, c_i$

Let $s_1 < \ldots < s_J$ be the ordered distinct observed event times.

$$\hat{S}_{KM}(t) = \prod_{t \geq s_j} \left(1 - \frac{d_j}{n_j}\right)$$

$n_j =\#$ at risk to an event at s_j: $\#\{i : u_i \geq s_j\}$
$d_j =\#$ observed event times at s_j: $\#\{i : u_i = s_j, \delta_i = 1\}$

Example. Freireich et al. (1963, Blood, 21:699:716) applied 6-Mercaptopurine and a placebo to 42 youths (younger than 20 years) with leukaemia. The times of interest are the duration of remission in weeks. The data from the treatment group are

6-MP: 6, 6, 6, 7, 10, 13, 16, 22, 23, 6+, 9+, 10+, 11+, 17+, 19+, 20+, 25+, 32+, 32+, 34+, 35+

(the notation t+ to indicate a right-censored observation at time t)

Estimated Cumulative Incidence by t: $1 - \hat{S}_{KM}(t)$

- CI: $\left(1 - \hat{S}_{KM}(t)\right) \pm 1.96SE$; $SE^2 = \hat{S}_{KM}(t)^2 \sum_{s_j \leq t} \frac{d_j}{n_j(n_j - d_j)}$

- alternative approaches?

Inference with right-censored event times?

Independent Censoring Assumption

- A key assumption in KM: censoring is unrelated to the event time

For another example, when $CI(t) = 1 - exp(-\lambda t)$, how to estimate IR with censored-data? $(IR = \lambda)$

Right-Censored Data $\{y_i = (u_i, \delta_i) : i = 1, \ldots n\}$ with independent censoring $C \sim g(\cdot)$: Let $\theta = 1/\lambda$

$$\begin{aligned} L(\theta|data) &= \prod_{i=1}^{n} \left[\theta^{-1} exp(-\frac{u_i}{\theta})\bar{G}(u_i)\right]^{\delta_i} \left[g(u_i) exp(-\frac{u_i}{\theta})\right]^{1-\delta_i} \\ &\propto \theta^{-\sum \delta_i} exp(-\sum u_i/\theta) \end{aligned}$$

$$\frac{\partial \log L(\theta|data)}{\partial \theta} = \frac{-\sum \delta_i}{\theta} + \frac{\sum u_i}{\theta^2} = 0$$

$\Longrightarrow$ MLE of θ: $\hat{\theta} = \sum_{i=1}^{n} u_i \big/ \sum_{i=1}^{n} \delta_i$

Part II.3.1 Overview of Designs for Medical Studies: Introduction

In general, a practical study is as follows

Problem Identification
(Specify question / hypothesis)
$\Downarrow$
Protocol Development
(Select a study design)
$\Downarrow$
Data Collection
$\Downarrow$
Data Analysis
$\Downarrow$
Conclusions

Part II.3.1 Introduction

Often conclusions are made via **descriptive analysis** and/or **statistical analysis**:

- Judgment is required to make inferences from the sampled population to the target population, *when the two populations are different*.
- Random sampling is required to make statistical inferences from the sample to the sampled population, *when the sample is a subset of the sampled population*.

Target Population
$\Downarrow \Uparrow$
Sampled Population
$\Downarrow \Uparrow$
Sample

$\Longrightarrow$ **Study design is very key.**

Part II.3.2 Types of Study Design

Example II.3.1. Data from a 1998 general social survey: a random sample of $n = 1127$ subjects were classified according to presence/absence of two characteristics, yes/no of belief in afterlife and female or male.

	Belief in Afterlife	
Gender	yes	no/undecided
female	509	116
male	398	104

What do you see from the data? What can be answered?
Before to do anything, how were the data collected and why?
$\Longrightarrow$ **study design?**

Part II.3.2. Types of Study Design

Example II.3.2. Data from a Harvard physicians' health study: enrolled subjects were randomized to placebo or aspirin group, and recorded were whether they had any heart attacks during the 5-year study.

	Myocardial Infarction	
Group	yes	no
placebo	189	10,845
aspirin	104	10,933

What do you see from the data? What can be answered?
Before to do anything, how were the data collected?
$\Longrightarrow$ **study design?**

Part II.3.2 Types of Study Design

- A **study design** is a plan for selecting study subjects and for obtaining data about them.
- There coulbe be many possible designs; but in practice, a few standard designs account for most epidemiologic research, say, and offer enough flexibility to address a wide range of research questions.
- There are many possible ways to classify study designs, depending on which features are highlighted.

Part II.3.2 Types of Study Design: Observational vs Experimental Studies

- An observational study draws inferences about the possible effect of a "treatment" on subjects, where the assignment of subjects into a "treated" group versus a "control" group is outside the control of the investigator.

  Example II.3.1. the social survey on belief in afterlife

- An experimental study, such as a randomized controlled trial, where each subject is randomly assigned to a "treated" group or a "control" group before the start of the treatment.

  Example II.3.2. the aspirin clinical trial

Part II.3.2. Types of Study Design: Cohort Study

- A **cohort** is a group of people who share a common characteristic or experience within a defined period (e.g., are born, are exposed to a drug or vaccine or pollutant, or undergo a certain medical procedure).
  Example II.3.1 is a cohort study

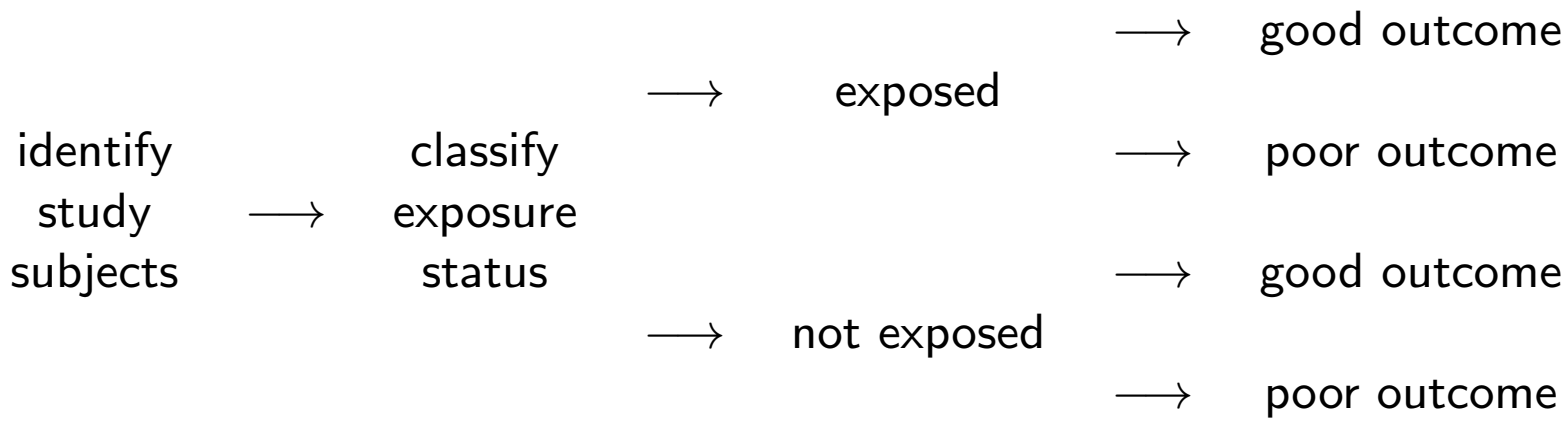


Figure 5-4. Cohort Study (Koepsell and Weiss, 2003)

Prospective, Retrospective, Ambidirectional Cohort Studies

Part II.3.2. Types of Study Design: Case-Control Study

Often cohort studies are not efficient with rare events/exposures such as cancer/smoking $\Longrightarrow$ another type of study design ...

Example II.3.3 An early study on the association of lung cancer with smoking: a random sample of 709 lung cancer patients, and a random sample of 709 non-lung cancer patients were respectively categorized according to ever smoked or not.

	Have Smoked	
Lung Cancer	yes	not
case	688	21
control	650	59

Part II.3.2. Types of Study Design: Case-Control Study

A **case-control** study compares the frequency of past exposure between *cases* who developed the disease and *controls* whom were chosen to reflect the frequency of exposure in the underlying population at risk from which the cases arose.

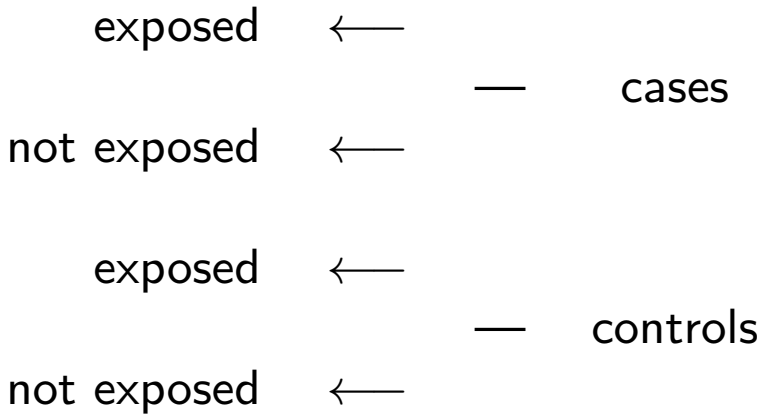


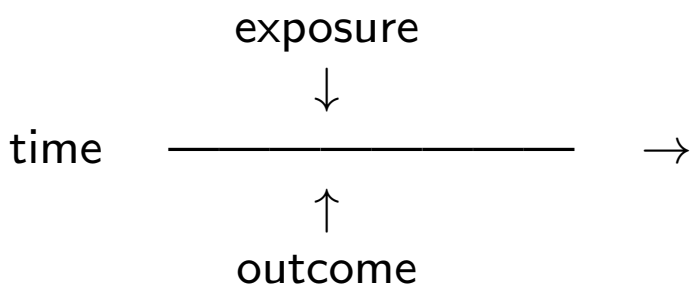
Figure 5-5. Case-Control Study (Koepsell and Weiss, 2003)

What a case-control study can/cannot answer?

Part II.3.2. Types of Study Design: Cross-Sectional Study

A **cross-sectional** study appears to ascertain the exposure and outcome at the same point/period in time.

Figure 5-7. Cross-Sectional Study (Koepsell and Weiss, 2003)



time trend? a cross-sectional clinical trial?

Part II.3.2. Types of Study Design: Longitudinal Study

A **longitudinal** study collects information on exposure/outcome repeatedly over time on study individuals: entitling the capacity of longitudinal studies to separate cohort and time effects

- contrasting case-control studies
- *repeated measures*, a classical version: record of a variable over time
- *time*, a generic term: the metameter for the occasions of observation
  comparing with *spatial data*

Part II.3.2. Types of Study Design: Randomized Trial

A **randomized trial** uses a formal chance mechanism to assign participants either to receive an intervention (or more) of interest or to serve as a control (or more).

- a solid basis for an inference of cause and effect
- *randomization* – control over confounding
- *experimental* vs *control* study arms – intervention vs placebo

more on clinical trial later ... ...
more on study design later ... ...

What to study next?

Part I – Introduction

Part II - Epidemiologic Concepts and Designs

- *II.1 Epidemiologic view of diseases and populations*
- *II.2 Measuring disease frequency in population*
- **II.3 Overview of designs for medical studies**

Part III - Clinical Trials

- **III.1 Clinical trial design principles**
- **III.2 Types of clinical trials**
- *III.3 Study monitoring*
- *III.4 A real-life example*

Part IV - Modern Biostatistical Methods